ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Linear Algebra and its Applications 430 (2009) 2581–2591
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Point estimation for multi-spectral distributed random matrices

Dong Q. Wang [a], S.E. Ahmed [b,*]

[a] School of Mathematics, Statistics and Computer Science, Victoria University of Wellington, P.O. Box 600, Wellington, New Zealand

[b] Department of Mathematics and Statistics, University of Windsor, Ontario, Canada N9B 3P4

Received 16 October 2007; accepted 19 June 2008
Available online 19 August 2008
Submitted by J.J. Hunter

Abstract

In this communication, we consider a $p \times n$ random matrix $X = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix}$ which is normally distributed with mean matrix M and covariance matrix Σ, where the multivariate observation $\mathbf{x}_i = \mathbf{y}_i + \boldsymbol{\epsilon}_i$ with p dimensions on an object consists of two components, the signal $\mathbf{y}_i$ with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ_s and noise $\boldsymbol{\epsilon}_i$ $(i = 1, 2, \ldots, n)$ with mean vector zero and covariance matrix Σ_ϵ, then the covariance matrix of $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by $\Sigma = \mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \Gamma \otimes (B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon)$, where Γ is a correlation matrix; $B_{|i-j|}$ and $C_{|i-j|}$ are diagonal constant matrices. The statistical objective is to consider the maximum likelihood estimate of the mean matrix M and various components of the covariance matrix Σ as well as their statistical properties, that is the point estimates of Σ_s, Σ_ϵ and Γ. More importantly, some properties of these estimators are investigated in slightly more general models.

Keywords: Covariance matrix and mean vector; Random matrices; Maximum likelihood estimation; Multivariate normal distribution; Noise and signal

* Corresponding author.
E-mail address: seahmed@uwindsor.ca (S.E. Ahmed).

0024-3795/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2008.06.034

1. Introduction

Let us define the $p \times n$ observation matrix X as

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix}. \tag{1.1}$$

Suppose that the $(p \times 1)$ observation vector $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ can be expressed as

$$\mathbf{x}_i = \mathbf{y}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{y}_i$ is assumed to have a p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ_s and the noise component $\boldsymbol{\epsilon}_i$ is also a multivariate normal with mean vector zero and covariance matrix Σ_ϵ. It is assumed further that $\mathbf{y}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated, and we consider the following covariance structure in remotely sensed data [13]:

$$\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j) = r_{ij} B_{|i-j|} \Sigma_s,$$

and

$$\mathrm{Cov}(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = r_{ij} C_{|i-j|} \Sigma_\epsilon, \tag{1.2}$$

where $|i - j|$ is an index, $i, j = 1, 2, \ldots, n$. $B_{|i-j|}$ and $C_{|i-j|}$ are diagonal constant matrices with the relationship that $B_{|i-j|} = (1 + kr_{ij})^{-1}[C_{|i-j|} + k \times I]$, and $|i - j| = 0, 1, 2, \ldots, n - 1$, where k is a known constant and $r_{ij} = r_{ji}$ are elements of the correlation matrix between random vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, and $r_{ss} = 1$. Note that if we let $B_0 = C_0 = I$, then the multivariate normal distribution model for a random matrix is obtained [14,2]. If we also assume $\Sigma = \Sigma_s + \Sigma_\epsilon$, then this model becomes the proportional covariance model [13,5, Section 3.1]. Hence this model is adopted here as a slightly generalized model or multi-spectral distribution model. In Green's paper (1988), the authors considered the problem of separating noise from the signal for some types of data, and developed a procedure for estimating some of the parameters.

The main results of this article are on the maximum likelihood estimation of parameters in a slightly more general model.

Suppose that the covariance matrix between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by

$$\Sigma = (\Sigma_{|i-j|}) = \mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \Gamma \otimes (B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon). \tag{1.3}$$

The elements of Γ are r_{ij} with $r_{ij} \in [-1, 1]$ and $M = \boldsymbol{\mu}\mathbf{1}^{\mathrm{T}}$, where $\mathbf{1}$ is a column vector of unit elements, then the probability density function of X can be written as

$$\begin{aligned} &f(X|M, \Sigma_s, \Sigma_\epsilon, \Gamma, B_{|i-j|}) \\ &\quad = (2\pi)^{-\frac{np}{2}} |B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon|^{-\frac{n}{2}} |\Gamma|^{-\frac{n}{2}} \\ &\quad\quad \times \exp\left\{-\frac{1}{2}\mathrm{tr}[(B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon)^{-1} \times (X - M)\Gamma^{-1}(X - M)^{\mathrm{T}}]\right\}, \end{aligned} \tag{1.4}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

The maximum likelihood estimators of M and Σ_s (for known $B_{|i-j|}$, Σ_ϵ and Γ) are respectively, given as follows:

$$\hat{M} = \hat{\boldsymbol{\mu}}\mathbf{1}^{\mathrm{T}} = (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1} X\Gamma\mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{1.5}$$

and

$$\hat{\Sigma}_s = n^{-1} B_{|i-j|}^{-1} (X - \hat{M}) \Gamma^{-1} (X - \hat{M})^{\mathrm{T}} - B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon. \tag{1.6}$$

These results follow directly from the results in [4]. Hence, we omit the details.

In subsequent sections, we consider the problem of estimating Γ and Σ_ϵ for a specific model. Note that the estimation of Σ_ϵ is only possible after the separation of the 'noise' from the 'signal'. In Section 2, we briefly describe two methods which have been used to separate noise from the signal. We use this 'signal–noise' terminology here, even if some of the models adopted may not be appropriate for the usual remotely sensed signals. In Section 3, the estimation of the correlation matrix Γ is considered. The estimation of Covariance matrix Σ_ϵ is dealt with in Section 4. Finally, some useful properties of these estimators are established in Section 5.

2. Separation of noise from signal

Principal components analysis has been traditionally used for the separation of noise from signal for some types of data, such as remotely sensed data. We refer to, Singh and Harrison [11], Ready and Wintz [8] Storvik [12] and Gillespie [1], among others. The question of how many principal components to retain (i.e., to represent the signal) can be determined using the cross-validation technique which was developed by [15,3]. Briefly, this method involves making a singular value decomposition of the observation matrix $X = \{x_{ij}\}$ with dimensions $p \times n$, such that X can be written as

$$X = \sum_{t=1}^{p} \mathbf{u}_t \lambda_t \mathbf{v}_t^{\mathrm{T}},$$

or

$$x_{ij} = \sum_{t=1}^{p} u_{it} \lambda_t v_{tj}, \tag{2.1}$$

where $i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, n$ and $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ are real square roots of the positive eigenvalues of the $p \times p$ matrix XX^{T}. The vectors $\mathbf{u}_t$ and $\mathbf{v}_t$ are eigenvectors of XX^{T} and $X^{\mathrm{T}}X$, respectively. Cross-validation methods can then be used to determine the number k of principal components which represent the 'signal' in the data, see [3]. Thus, the elements of X can be rewritten as

$$x_{ij} = \sum_{t=1}^{k} u_{it} \lambda_t v_{tj} + \varepsilon_{ij}, \tag{2.2}$$

where ε_{ij} represents the 'noise' component, and $\sum_{t=1}^{k} u_{it} \lambda_t v_{tj}$ represents the signal.

Interestingly, Green et al. [2] discussed some of the drawbacks associated with using principal components to separate noise from signal. Further, they proposed a transformation with the specific objective of separating noise and signal. In particular, under the model that the observation matrix X can be decomposed into independent signal and noise components, the observation vector $\mathbf{x}_i$ is expressed as $\mathbf{x}_i = \mathbf{y}_i + \boldsymbol{\epsilon}_i$. In this case, the covariance matrices can be written as

$$\Sigma^* = \Sigma_s + \Sigma_\epsilon, \tag{2.3}$$

where Σ^* is the covariance matrix of $\mathbf{x}_i$. The transformation developed by Green et al. was called the maximum noise fraction (MNF) transformation. In theory the procedure developed using

MNF improves the quality of the image by removing the noise component which is uncorrelated with the signal [13]. Having said that, in practice, this involves obtaining eigenvalues of the matrix $\Sigma_{|i-j|}(\Sigma^{*})^{-1}$.

In the above expression, pixel or object difference $|i - j|$ would be appropriate for estimation of noise. Clearly, one needs to estimate these parameters in practice and this is no trivial task for some images, as is discussed in [7].

In subsequent sections, we assume that it is possible to separate signal from noise, and we present results under the models as defined in Section 1.

3. Estimation of correlation matrix Γ

Wang and Lawoko [14] obtained results regarding the maximum likelihood estimation of Γ for a normally distributed population and the same method is used here. Consider the function given in (1.4), which can be rewritten as

$$\begin{aligned} L &= \log f(M, \Sigma_\epsilon, \Sigma_s, \Gamma|X, B_{|i-j|}) \\ &= \text{constant} + \frac{n}{2}\log|\Delta^{-1}| + \frac{p}{2}\log|\Gamma^{-1}| \\ &\quad -\frac{1}{2}\mathrm{tr}[H\Delta^{-1}H^{\mathrm{T}}(X-M)\Gamma^{-1}(X-M)^{\mathrm{T}}], \end{aligned}$$

where $\Delta = H^{\mathrm{T}}(B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon)H$, and H is an orthogonal matrix.

The differentiation of L with respect to Γ and Δ yield

$$\begin{aligned} \mathrm{d}L &= \left\{\frac{n}{2}\mathrm{tr}(\Delta) - \frac{1}{2}\mathrm{tr}[H(X-M)\Gamma^{-1}(X-M)^{\mathrm{T}}H^{\mathrm{T}}]\right\}\mathrm{d}\Delta^{-1} \\ &\quad + \left\{\frac{n}{2}\mathrm{tr}(\Gamma) - \frac{1}{2}\mathrm{tr}[(X-M)H\Delta^{-1}H^{\mathrm{T}}(X-M)^{\mathrm{T}}]\right\}\mathrm{d}\Gamma^{-1}. \end{aligned}$$

Equating $\mathrm{d}L$ to zero, we obtain the following equations.

$$\begin{aligned} &(X-M)\Gamma^{-1}(X-M)^{\mathrm{T}} = n(B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon) \\ &(X-M)(B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon)^{-1}(X-M)^{\mathrm{T}} = p\Gamma. \end{aligned}$$

Noting that, the $p \times p$ matrix is non-singular and $M = (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}X\Gamma^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}$, then we obtain the following matrix equation as a function of Γ^{-1}.

$$\begin{aligned} &XX^{\mathrm{T}} - (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}XX^{\mathrm{T}}\Gamma^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}} - (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}XX^{\mathrm{T}} \\ &\quad + (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-2}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}XX^{\mathrm{T}}\Gamma^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}} = O_{p\times p}, \end{aligned}$$

where $O_{p\times p}$ denotes a zero matrix of dimension $p \times p$, and $\mathbf{1}$ is a column vector of unit elements. Then the equation

$$[I - (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}]XX^{\mathrm{T}}[I - (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}]^{\mathrm{T}} = O_{p\times p},$$

which is a quadratic form in $I - (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}$ can be solved numerically. Thus, the elements of $\hat{\Gamma}$ can be readily obtained via numerical computations. Further, adjustments are required on the elements of $\hat{\Gamma}$ in order to get a 'true' correlation matrix. Details of the methodology is available in [9].

Note, if the matrix Γ is non-singular, then we consider the eigenvalues of Γ. Let the non-zero eigenvalues be $\lambda_1, \lambda_2, \ldots, \lambda_p$, and the matrix containing the corresponding set of normalized eigenvectors, V, then we can write $\Gamma^{-1} = VWV^{\mathrm{T}}$, where W is a diagonal matrix whose jth diagonal element is

$$W_j = \frac{1}{\max\{\lambda_j, \lambda_s\}}, \quad j = 1, 2, \ldots, n,$$

and λ_s is the smallest non-zero eigenvalue of Γ.

4. Estimation of covariance matrix Σ_ϵ

If the signal and noise components of the observation vectors can be separated by the methods suggested in Section 2, and the observation vector can be written as $\mathbf{x}_i = \mathbf{y}_i + \epsilon_i$, $i = 1, 2, \ldots, n$, where

$$\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{Cov}(\mathbf{y}_i + \epsilon_i, \mathbf{y}_j + \epsilon_j) = r_{ij} B_{|i-j|} \Sigma_s + r_{ij} C_{|i-j|} \Sigma_\epsilon,$$

where $B_{|i-j|}$ (or $C_{|i-j|}$) and $r_{ij} = \hat{r}_{ij}$ are known, then it can be established that

$$\begin{aligned} \Sigma_{|i-j|} &= \mathrm{Cov}(\mathbf{x}_i - \mathbf{x}_j) \\ &= E[\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}] - E[\mathbf{x}_i \mathbf{x}_j^{\mathrm{T}}] - E[\mathbf{x}_j \mathbf{x}_i^{\mathrm{T}}] + E[\mathbf{x}_j \mathbf{x}_j^{\mathrm{T}}] \\ &= 2[\Sigma - \mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)] \\ &= 2[(I - r_{ij} B_{|i-j|})\Sigma^* + r_{ij}(B_{|i-j|} - C_{|i-j|})\Sigma_\epsilon] \\ &= 2(I - r_{ij} B_{|i-j|})(\Sigma^* + k r_{ij} \Sigma_\epsilon), \end{aligned}$$

where $\Sigma^* = \Sigma_s + \Sigma_\epsilon$.

Now consider

$$\Sigma_{|i-j|}(\Sigma^*)^{-1} = 2(I - r_{ij} B_{|i-j|})(\Sigma^* + k r_{ij} \Sigma_\epsilon)(\Sigma^*)^{-1},$$

so that

$$(I - r_{ij} B_{|i-j|})^{-1} \Sigma_{|i-j|}(\Sigma^*)^{-1} = 2(I + k r_{ij} \Sigma_\epsilon (\Sigma^*)^{-1}),$$

where $\Sigma_\epsilon(\Sigma^*)^{-1}$ is a positive definite matrix. Non-singular matrices H and G can be found respectively so that, [10]

$$H \Sigma_\epsilon (\Sigma^*)^{-1} G = \Lambda,$$

where Λ is a diagonal matrix. Thus,

$$(I - r_{ij} B_{|i-j|})^{-1} \Sigma_{|i-j|} \Sigma^{-1} = 2H^{-1}(HG + k r_{ij} \Lambda)G^{-1} = H^{-1} \Lambda^* G^{-1},$$

where $\Lambda^* = 2(HG + k r_{ij} \Lambda)$.

This shows that $(I - r_{ij} B_{|i-j|})^{-1} \Sigma_{|i-j|}(\Sigma^*)^{-1}$ has the same (normalized) eigenvectors as $\Sigma_\epsilon(\Sigma^*)^{-1}$, and the estimation of Σ_ϵ is transformed into calculating the eigenvalues Λ^* and eigenvectors H of the matrix $(I - r_{ij} B_{|i-j|})^{-1} \Sigma_{|i-j|}(\Sigma^*)^{-1}$.

Hence one can find H and G for a certain $|i - j|$ and $(i \neq j)$, then substitute Λ from

$$\Lambda = (2k r_{ij})^{-1}(\Lambda^* - 2HG)$$

into $\Sigma_\epsilon = H^{-1}\Lambda G^{-1}\Sigma^*$, and an estimate of Σ_ϵ is obtained as

$$\hat{\Sigma}_\epsilon = (2kr_{ij})^{-1}H^{-1}(\Lambda^* - 2HG)G^{-1}\hat{\Sigma}^*,$$

where $\hat{\Sigma}^*$ is obtained from the observation matrix X. Note that this method will give a different value of $\hat{\Sigma}_\epsilon$ for different values of $|i-j|$. In practice, one would use $|i-j| = 1$ or 2.

Finally, we consider the estimation of M and Σ_s. Noting that, for known $B_{|i-j|}$, Γ and Σ_ϵ, the maximum likelihood estimators of M and Σ_s along with some of their statistical properties were given in [14]. If the noise can be separated from the signal then the results of this article can be used to estimate Γ and Σ_ϵ. Consequently, these estimates can be inserted in the expressions (1.5) and (1.6) respectively, for $\hat{M}$ and $\hat{\Sigma}_s$ to get the estimates of M and Σ_s for unknown Γ and Σ_ϵ.

5. Statistical properties of estimators

Suppose now that $C_{|i-j|}$ (hence $B_{|i-j|}$), Σ_ϵ and Γ are known and attention is focused on the estimation of M and Σ_s. Under these conditions the following statistical properties of $\hat{M}$ and $\hat{\Sigma}_s$ are modifications of certain results in [14]. Consequently, only brief proofs are presented here (apart from Result 5 which requires more substantial derivation).

Theorem 1. *The maximum likelihood estimators of M and Σ_s are those given in expressions* (1.5) *and* (1.6).

Proof. The proof is identical to that of a similar result in [4]. □

Theorem 2. *The maximum likelihood estimator $\hat{M}$ of M is an unbiased estimator, and its covariance is given by*

$$\mathrm{Cov}(\hat{M}) = n(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}B_{|i-j|}\Sigma_s + n(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}C_{|i-j|}\Sigma_\epsilon.$$

Proof. From (1.5) we have

$$\begin{aligned} E(\hat{M}) &= E[(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}X\Gamma^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}] \\ &= (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}\boldsymbol{\mu}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}} \\ &= \boldsymbol{\mu}\mathbf{1}^{\mathrm{T}} = M. \end{aligned}$$

To derive the covariance of $\hat{M}$, we consider

$$\begin{aligned} E[(\hat{M}-M)(\hat{M}-M)^{\mathrm{T}}] &= (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-2}E[(X-M)\Gamma^{-1}(\mathbf{1}\mathbf{1}^{\mathrm{T}})^2\Gamma^{-1}(X-M)^{\mathrm{T}}] \\ &= (\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{2}\mathrm{tr}[(\Gamma^{-1}(\mathbf{1}\mathbf{1}^{\mathrm{T}})^2\Gamma^{-1})](B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon) \\ &= n(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}(B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon). \end{aligned}$$

Therefore,

$$\mathrm{Cov}(\hat{M}) = n(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}B_{|i-j|}\Sigma_s + n(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}C_{|i-j|}\Sigma_\epsilon. \qquad \square$$

Theorem 3. *The maximum likelihood estimator $\hat{\Sigma}_s$ of Σ_s is a biased estimator, and*

$$E(\hat{\Sigma}_s) = \frac{n-1}{n}\Sigma_s - \frac{1}{n}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon.$$

Proof. From expression (1.6), we have

$$E(\hat{\Sigma}_s) = n^{-1}B_{|i-j|}^{-1}E[(X-\hat{M})\Gamma^{-1}(X-\hat{M})^{\mathrm{T}}] - B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon$$
$$= n^{-1}B_{|i-j|}^{-1}E\{X\Gamma^{-1}[I-(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}]X^{\mathrm{T}}\} - B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon.$$

After some algebraic manipulations, we obtain

$$E(\hat{\Sigma}_s) = n^{-1}(n-1)B_{|i-j|}^{-1}(B_{|i-j|}\Sigma_s + C_{|i-j|}\Sigma_\epsilon) - B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon$$
$$= \frac{n-1}{n}\Sigma_s + \left(\frac{n-1}{n}-1\right)B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon$$
$$= \frac{n-1}{n}\Sigma_s - \frac{1}{n}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon. \qquad \square$$

Theorem 4. *The unbiased maximum likelihood estimator of Σ_s is*

$$n(n-1)^{-1}\hat{\Sigma}_s + (n-1)^{-1}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon.$$

Proof. Consider

$$E[n(n-1)^{-1}\hat{\Sigma}_s + (n-1)^{-1}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon]$$
$$= n(n-1)^{-1}E[\hat{\Sigma}_s] + (n-1)^{-1}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon$$
$$= n(n-1)^{-1}\left[\frac{n-1}{n}\Sigma_s - \frac{1}{n}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon\right] + (n-1)^{-1}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon$$
$$= \Sigma_s. \qquad \square$$

Theorem 5. *Let*

$$\hat{\Sigma}_s^* = n(n-1)^{-1}\hat{\Sigma}_s + (n-1)^{-1}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon.$$

The matrix $\hat{\Sigma}_s^*$ is an unbiased estimator of Σ_s, and the covariance of $\hat{\Sigma}_s^*$ is given by

$$\mathrm{Cov}(\hat{\Sigma}_s^*) = \frac{1}{n-1}\{n\Sigma_{|i-j|}^2 + \Sigma_{|i-j|}[\mathrm{tr}(\Sigma_{|i-j|}^2)I - 2(n-1)\mathrm{tr}(D^*\Gamma)B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon]\}$$
$$+\Sigma_\epsilon^{\mathrm{T}}B_{|i-j|}^2C_{|i-j|}^2\Sigma_\epsilon - \Sigma_s - \Sigma_\epsilon,$$

where $\Sigma_{|i-j|} = \mathrm{Cov}(\mathbf{x}_i - \mathbf{x}_j)$.

Proof. From result 4, we know that $E(\hat{\Sigma}_s^*) = \Sigma_s$; thus

$$\mathrm{Cov}(\hat{\Sigma}_s^*) = E[(\hat{\Sigma}_s^*)^2] - \Sigma_s^2.$$

Using expression (1.5) we have

$$(X-\hat{M})\Gamma^{-1}(X-\hat{M})^{\mathrm{T}} = XDX^{\mathrm{T}},$$

where $D = \Gamma^{-1}[I-(\mathbf{1}^{\mathrm{T}}\Gamma^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Gamma^{-1}]$. From expression (1.6) and the above results it follows that

$$\hat{\Sigma}_s^* = n(n-1)^{-1}[n^{-1}B_{|i-j|}^{-1}(X-\hat{M})\Gamma^{-1}(X-\hat{M})^{\mathrm{T}} - B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon]$$
$$+(n-1)^{-1}B_{|i-j|}^{-1}C_{|i-j|}\Sigma_\epsilon$$

$$
\begin{aligned}
&= (n-1)^{-1} B_{|i-j|}^{-1} X D X^{\mathrm{T}} - n(n-1)^{-1} B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon \\
&\quad + (n-1)^{-1} B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon \\
&\quad = X D^* X^{\mathrm{T}} - B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon,
\end{aligned}
$$

where $D^* = (n-1)^{-1} B_{|i-j|}^{-1} D$. Thus,

$$
\begin{aligned}
(\hat{\Sigma}_s^*)^2 &= [X D^* X^{\mathrm{T}} - B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon]^2 \\
&= X D^* X^{\mathrm{T}} X D^* X^{\mathrm{T}} - 2 X D^* X^{\mathrm{T}} B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon + \Sigma_\epsilon^{\mathrm{T}} B_{|i-j|}^{-2} C_{|i-j|}^2 \Sigma_\epsilon.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
E(\hat{\Sigma}_s^*)^2 &= E(X D^* X^{\mathrm{T}} X D^* X^{\mathrm{T}}) - 2E(X D^* X^{\mathrm{T}}) B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon \\
&\quad + \Sigma_\epsilon^{\mathrm{T}} B_{|i-j|}^{-2} C_{|i-j|}^2 \Sigma_\epsilon.
\end{aligned} \tag{5.1}
$$

Using results in [6,14], the following expressions are obtained:

$$
\begin{aligned}
&E(X D^* X^{\mathrm{T}} X D^* X^{\mathrm{T}}) \\
&\quad = [1 + (n-1)^{-1}](B_{|i-j|} \Sigma_s + C_{|i-j|} \Sigma_\epsilon)^2 \\
&\quad\quad + (n-1)^{-1} [\mathrm{tr}(B_{|i-j|} \Sigma_s + C_{|i-j|} \Sigma_\epsilon)] \times (B_{|i-j|} \Sigma_s + C_{|i-j|} \Sigma_\epsilon).
\end{aligned} \tag{5.2}
$$

and

$$
E(X D^* X^{\mathrm{T}}) = \mathrm{tr}(D^* \Gamma)(B_{|i-j|} \Sigma_s + C_{|i-j|} \Sigma_\epsilon).
$$

Using expressions (5.2) and the above equation, the expression (5.1) is rewritten after some matrix algebra as

$$
\begin{aligned}
E(\hat{\Sigma}_s^*)^2 &= \frac{n}{n-1} [\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)]^2 + \frac{1}{n-1} \mathrm{tr}[\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)] \mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) \\
&\quad - 2\mathrm{tr}(D^* \Gamma) \mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j) B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon + \Sigma_\epsilon^{\mathrm{T}} B_{|i-j|}^{2} C_{|i-j|}^2 \Sigma_\epsilon.
\end{aligned}
$$

Since $\Sigma_{|i-j|} = \mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)$, we have

$$
\begin{aligned}
\mathrm{Cov}(\hat{\Sigma}_s^*) &= \frac{1}{n} \{ n \Sigma_{|i-j|}^2 + \Sigma_{|i-j|} [\mathrm{tr}(\Sigma_{|i-j|}^2) I - 2(n-1) \mathrm{tr}(D^* \Gamma) B_{|i-j|}^{-1} C_{|i-j|} \Sigma_\epsilon] \} \\
&\quad + \Sigma_\epsilon^{\mathrm{T}} B_{|i-j|}^{2} C_{|i-j|}^2 \Sigma_\epsilon - \Sigma^*,
\end{aligned}
$$

which yields the required result. □

6. An application

Given $B_{|i-j|} = C_{|i-j|} = I$, when Σ_ϵ is known and Σ_s, Γ are estimated as in Sections 3 and 4, the covariance matrix $\hat{\Gamma} \otimes (\hat{\Sigma}_s + \Sigma_\epsilon)$ (support random matrix) of X is obtained by maximum likelihood method. Then the modified discriminant function (MDF) for each class j is constructed by

Fig. 1. Flipped digit examples.

$$F_j(\mathbf{x}) = -\ln|\hat{\Gamma} \otimes (\hat{\Sigma}_s + \Sigma_\epsilon)| - (\mathbf{x} - \bar{\mathbf{x}}_j)^{\mathrm{T}}(\hat{\Gamma} \otimes (\hat{\Sigma}_s + \Sigma_\epsilon))^{-1}(\mathbf{x} - \bar{\mathbf{x}}_j),$$

where $j = 1, 2, \ldots, c$; $\bar{\mathbf{x}}_j$ is an estimator of the mean vector of the jth cluster of the random matrix X, and c is the number of classes. Note that we use the covariances $\hat{\Gamma} \otimes (\hat{\Sigma}_s + \Sigma_\epsilon)$ in MDF. The classification rule is given as follows: If for any given observation, $\mathbf{x}^*$, $F_i(\mathbf{x}^*) \geqslant F_j(\mathbf{x})$, for all $j \neq i$, then the item $\mathbf{x}^*$ is assigned to class i.

A simulation data set is generated to investigate the power of our new approach. We used nine multiple dimension digit recognition tasks in the experiments. Each task involves a file (a collection) of digit images. Each file contains 100 examples for each of the 10 digits (0, 1, ..., 9), making a total number of 1000 digit examples. Each digit example is an image of a 7×7 bitmap. These tasks were chosen to provide classification problems of increasing difficulty, as shown in Table 1. In all of these recognition problems, the goal is to automatically recognize which of the 10 classes (digits 0, 1, 2, ..., 9) each pattern belongs to. Except for the first file which contains clean patterns, all data patterns in the other eight files have been corrupted by noise. The amount of noise in different files was randomly generated based on the percentage of flipped pixels and was given by the two numbers `nn` in the file name. For example, the first row of this table shows that recognition task 1 is to classify those clean digit patterns into the ten different classes. In this task, there are 1000 patterns in total, 500 are used for training and 500 for testing. In task 3, however, 10% of pixels, chosen at random, have been flipped. All the training examples are randomly ordered.

Examples of the 9 tasks are shown in Fig. 1. The 9 lines of digit samples correspond to the 9 recognition tasks in Table 1. The first 3 tasks, one with clean data and two with only 5% and 10% of flipped rate, are relatively straightforward for human eyes, though there is still some difficulty in distinguishing between “3” and “9”. With the increase in the flipped rate in these patterns such as task 4 and task 5, it becomes more difficult to classify these digit patterns, even if humans can

Table 1
Nine digit recognition tasks

Task	File name	Noise amount	Total patterns	Training set	Test set
1	digit00	0%	1000	500	500
2	digit05	5%	1000	500	500
3	digit10	10%	1000	500	500
4	digit15	15%	1000	500	500
5	digit20	20%	1000	500	500
6	digit30	30%	1000	500	500
7	digit40	40%	1000	500	500
8	digit50	50%	1000	500	500
9	digit60	60%	1000	500	500

Table 2
Results for optimal error rate (OER) for task 6

	Digit									
	0	1	2	3	4	5	6	7	8	9
OER	0.00	0.00	0.00	0.25	0.00	0.09	0.16	0.00	0.19	0.25
Prior probability	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10

still recognize the majority. From task 6 to task 9, however, it is very difficult, even impossible, for human eyes to make the discrimination. We hypothesized that our new method will do a good job for the first three tasks, but can not be excellent for tasks 6–9. We also want to investigate whether our new method can achieve an acceptable performance for these difficult tasks and whether the new method outperforms neural networks on these tasks.

This example is also used to briefly describe how to obtain the classification error for each task by applying the MDF. After applying MDF to each task, the classes of discriminant function can be obtained. For example, there is 30% of noise flipped in the 1000 digits in task 6. The classification results for test data are summarized in Table 2. The total optimal error rate (OER) for this task is 0.106, or the classification accuracy is 89.40% on average of 10 runs.

Acknowledgments

We would like to thank the referee and co-editor for very constructive comments and suggestions on earlier version of this article. The research was supported by grants from URF, Victoria University, New Zealand, and Natural Sciences and Engineering Research Council of Canada.

References

[1] A.R. Gillespie, Digital techniques of image enhancement, in: B.S. Siegal, A.R. Gillespie (Eds.), Remote Sensing in Geology, Wiley, New York, 1980, pp. 195–202 and 220–203.
[2] A.A. Green, M. Berman, P. Switzer, M.D. Craig, A transformation for ordering multi-spectral data in terms of image quality with implications for noise removal, IEEE Trans. Geosci. Remote Sensing 26 (1) (1988) 65–74.
[3] W.J. Krzanowski, Cross-validatory choice in principal component analysis some sampling results, J. Statist. Comput. Simul. 8 (1983) 299–314.
[4] C.R.O. Lawoko, On the maximum likelihood estimation of parameters of a spatial discrimination model, Commun. Statist. – Theory Methods 19 (12) (1990) 4627–4641.
[5] J.R. Muirhead, Aspects of Multivariate Statistical Theory, John Wiley, New York, 1982.
[6] H. Neudecker, T. Wansbeek, Fourth order properties of normally distributed randoms matrices, Linear Algebra Appl. 97 (1987) 13–21.
[7] N. Pendock, A.A. Nielsen, Multi-spectral image enhancement neural networks and the maximum noise fraction transform, in: Proceedings of the 7th Australasian Remote Sensing Conference (Melbourne), 1994, pp. 415–422.
[8] P.J. Ready, A.P. Wintz, Information extraction, SNR improvement, and data compression in multi-spectral imagery, IEEE Trans. Commun. 21 (1973) 1123–1130.
[9] P.J. Rousseeuw, G. Molenberghs, Transformation of nonpositive semidefinite correlation matrices, Commun. Statist. – Theory Methods 22 (4) (1993) 965–984.
[10] C.R. Rao, Linear Statistical Inference and its Applications, John Wiley and Sons Inc., 1973, pp. 19–20.
[11] A. Singh, A. Harrison, Standardized principal components, Int. J. Remote Sensing 6 (1985) 883–896.
[12] G. Storvik, Data-reduction by projection for multivariate spatial images, Research Report, 1994, ISBN 82-539-0346-4.

[13] P. Switzer, Min/Max autocorrelation factors for multivariate spatial imagery, in: L. Billard (Ed.), Computer Science and Statistics: The Interface, North Holland, Amsterdam, 1980, pp. 13–16.
[14] D.Q. Wang, C.R.O. Lawoko, Estimation of parameters for normally distributed random matrices, Linear Algebra Appl. 210 (1995) 199–208.
[15] S. Wold, Cross-validatory estimation of the number of components in factor and principal component models, Technometrics 20 (1978) 397–405.